

Efficient Reverse k Nearest Neighbor evaluation for hierarchical index

Siddharth Dawar, Indraprastha Institute of Information Technology, India

Vikram Goyal, Indraprastha Institute of Information Technology, India

Debajyoti Bera, Indraprastha Institute of Information Technology, India

“Reverse Nearest Neighbor” query finds applications in decision support systems, profile-based marketing, emergency services etc. In this paper, we point out a few flaws in the branch and bound algorithms proposed earlier for computing monochromatic $RkNN$ queries over data points stored in hierarchical index. We give suitable counter examples to validate our claims and propose a correct algorithm for the corresponding problem. We show that our algorithm is correct by identifying necessary conditions behind correctness of algorithms for this problem.

1. INTRODUCTION

One important type of operation that is gaining popularity in database and data-mining research community is the *Reverse Nearest Neighbor Query* ($RkNN$) [Korn and Muthukrishnan 2000]. Given a set of database objects O and a query object Q , the $RkNN$ query returns those objects in O , for which Q is one of their k nearest neighbors; here the notion of neighborhood is with respect to an appropriately defined notion of distance between the objects. A classic example $RkNN$ is in the domain of decision support systems where the task is to open a new facility (like a restaurant) in an area such that it will be least influenced by its competitors and attract good business. Another application is profile based marketing [Korn and Muthukrishnan 2000], where a company maintains profiles of its customers and wants to start a new service which can attract the maximum number of customers. $RkNN$ has also applications in clustering, where a cluster could be created by identifying a group of objects, and clustering them around their common nearest neighbor point – this essentially involves finding cluster centers with high cardinality of reverse nearest neighbor sets. *Reciprocal nearest neighborhood*, in which data points which are nearest neighbors of each other are clustered together (and therefore, satisfy both nearest neighbor and reverse nearest neighbor criteria), is another well-known technique in clustering [López-Sastre et al. 2012].

This important concept has seen a series of remarkable applications and algorithms for processing different types of objects, in various contexts and under variations [Kang et al. 2007], [Safar et al. 2009], [Tran et al. 2009], [Taniar et al. 2011], [Shang et al. 2011], [Cheema et al. 2012], [Ghaemi et al. 2012], [Li et al. 2013], [Emrich et al. 2014], [Cabello et al. 2010], [Bhattacharya and Nandy 2013] of the problem parameters. The focus of this paper is monochromatic $RkNN$ queries – in this version, all objects in the database and the query belong to the same category, unlike the bichromatic version in which the objects can belong to different categories. Furthermore, we want to focus on queries where k is specified as part of a query, and want to support objects from an arbitrary metric space.

This paper points out several fundamental inaccuracies in three papers published earlier on the problem mentioned above.

- Reverse k -nearest neighbor search in dynamic and general metric databases [Achtert et al. 2009]
- Reverse spatial and textual k nearest neighbor search [Lu et al. 2011]
- Efficient algorithms and cost models for reverse spatial-keyword k -nearest neighbor search [Lu et al. 2014]

Achtert et al. [Achtert et al. 2009] proposed a branch-and-bound algorithm for the above problem which could use any given hierarchical tree-like index on data from any metric space. Lu et al. [Lu et al. 2011] proposed a similar algorithm, but specifically optimized for spatio-textual data, for answering RST k NN queries using a specialized IUR tree as the indexing structure. In a followup paper [Lu et al. 2014], they proposed an improvement of their algorithm (including correcting an error) and a theoretical cost model to analyze the efficiency of their algorithm. However, we observed several deficiencies in the algorithms mentioned above. In this paper we will point out those inaccuracies, and discuss them more formally by pointing out some key properties which these algorithms violate, but are necessary for ensuring correctness of these and other similar algorithms. We will present detailed counter examples and suggest corrective modifications to these algorithms. Finally we will propose a correct algorithm for performing R k NN queries over a hierarchical index and also present its proof of correctness.

The paper is organized as follows. In Section 2 we explain the three published approaches mentioned above in which we found inaccuracies. In Section 3 we describe our counter-examples with respect to them. We present our modified algorithm in Section 4, and its proof of correctness in Section 4.6.

2. EARLIER RESULTS

The underlying algorithms for all three approaches mentioned above essentially have the same structure and follow a branch-and-bound approach. The former work is applicable on any kind of data with a distance measure that is a metric, and uses any hierarchical tree-like index built on the data. The two latter work are specifically concerned with R k NN query on spatio-textual data, which they refer to as RST k NN query.

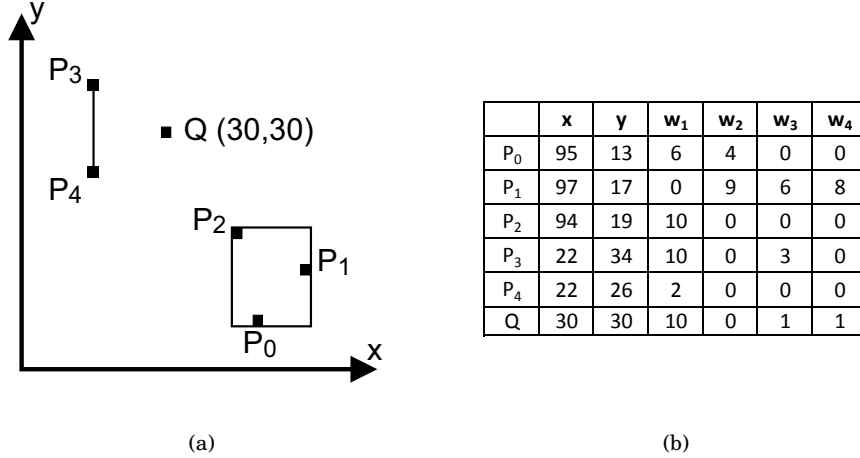


Fig. 1: Example for illustrating RST k NN and R k NN

In RST k NN, each object is represented by a pair (loc, vct) where loc is the spatial location and vct is the associated textual description which is represented by (word, weight(word)) pairs for all words appearing in the database. Weight of a word is calculated on the basis of TF-IDF scheme [Salton and Buckley 1988]. Spatio-textual

similarity ($SimST$) is defined by [Lu et al. 2011] as follows:

$$SimST(o_1, o_2) = \alpha * \left(1 - \frac{dist(o_1.loc, o_2.loc) - \varphi_s}{\psi_s - \varphi_s}\right) + (1 - \alpha) * \left(\frac{EJ(o_1.vct, o_2.vct) - \varphi_t}{\psi_t - \varphi_t}\right) \quad (1)$$

The parameter α is used to define the relevance factor for spatial and textual similarity while calculating the total similarity scores and is specified in a query. φ_s and ψ_s denote the minimum and maximum distance between any two objects in the database and are used to normalize the spatial similarity to the range $[0, 1]$. Similarly φ_t and ψ_t denote the minimum and maximum textual similarity between any two objects in the database. $dist(\cdot)$ is the Euclidean Distance between o_1 and o_2 and EJ is the Extended Jaccard Similarity [Tan and Steinbach 2011] defined as:

$$EJ(o_1.vct, o_2.vct) = \frac{\sum_{j=1}^n o_1.w_j * o_2.w'_j}{\sum_{j=1}^n o_1.w_j^2 + \sum_{j=1}^n o_2.w'_j^2 - \sum_{j=1}^n o_1.w_j * o_2.w'_j} \quad (2)$$

where $o_1.vct = \langle w_1, \dots, w_n \rangle$ and $o_2.vct = \langle w'_1, \dots, w'_n \rangle$.

As an example, consider Figure 1. There, considering only location attributes, and for $k = 2$, $RkNN$ of Q are objects P_3 and P_4 . However, if we consider both spatial and textual similarity, and taking $k = 2$ and $\alpha = 0.4$, $RSTkNN$ of Q is P_2, P_3 and P_4 .

Now we will describe the actual algorithm proposed by [Lu et al. 2011] for $RSTkNN$. It is important to present it in some detail – this is required for proper appreciation of the inaccuracies in this algorithm. This algorithm requires its data to be organized as an hierarchical index called as IUR-tree. IUR-Tree is a R-Tree [Guttman 1984]; where every node of the tree is embedded with Intersection and Union Vectors. The textual vectors contain the weight of every distinct item in the documents contained in the node. The weight of every item in the Intersection Vector (resp. Union Vector) is the minimum weight (resp. maximum weight) of all the items present in the documents contained in the node. During the execution of the algorithm, a lower and upper nearest-neighbor list/contribution list is created and maintained for each node in the IUR-Tree. The lower (resp. upper) contribution list stores the minimum (resp. maximum) similarity between the node and its neighbors.

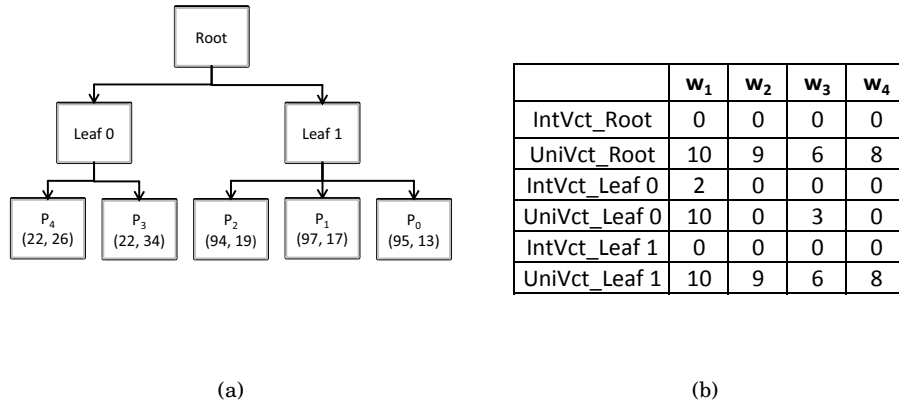


Fig. 2: IUR-Tree and Textual Vectors of Fig 1

Algorithm 1 RST k NN (R : IUR-Tree root, Q : query) from [Lu et al. 2011]

```

1: Output: All objects  $o$ , s.t  $o \in \text{RST}_k\text{NN}(Q, k, R)$ .
2: Initialize a priority queue  $U$ , and lists  $COL, ROL, PEL$ ;
3: EnQueue( $U, R$ );
4: while  $U$  is not empty do
5:    $P \leftarrow \text{DeQueue}(U)$ ; //Priority of  $U$  is  $\text{MaxST}(P, Q)$ 
6:   for each child node  $E$  of  $P$  do
7:     Inherit( $E.CLS, P.CLS$ );
8:     if IsHitOrDrop( $E, Q$ )==false then
9:       for each node  $E'$  in  $COL, ROL, U$  do //see subsection 3.2
10:        UpdateCL( $E, E'$ ); //update contribution lists of  $E$ ;
11:        if IsHitOrDrop( $E, Q$ )=true then //see subsection 3.3
12:          break;
13:        end if
14:        if  $E' \in U \cup COL$  then
15:          UpdateCL( $E', E$ ); //Update contribution Lists of  $E'$  using  $E$ .
16:          if IsHitOrDrop( $E', Q$ )==true then
17:            Remove  $E'$  from  $U$  or  $COL$ ;
18:          end if
19:        end if
20:        if  $E$  is not a hit or drop then
21:          if  $E$  is an index node then
22:            EnQueue( $U, E$ );
23:          else
24:             $COL.append(E)$ ; //a database object
25:          end if
26:        end if
27:      end for
28:    end if
29:  end for
30: end while
31: Final_Verification( $COL, PEL, Q$ );

```

The IUR-Tree and Intersection and Union Vectors of the corresponding nodes is shown in the Figure 2. These vectors along with the MBR's of nodes are used to compute the similarity approximations i.e. upper and lower bounds on the spatio-textual similarity between two groups of objects.

We refer to an internal node or a point in the IUR-Tree as an entry. The algorithm takes as an input an IUR-Tree (Intersection Union tree) R , query Q and returns all database objects which are RST k NN of Q . The data structures used are: a priority queue (U) sorted in decreasing order on $\text{MaxST}(E, Q)$, result list (ROL), pruned list (PEL) and candidate list (COL). $\text{MaxST}(E, Q)$ is the maximum spatial textual similarity of the entry E with the query point Q . The algorithm dequeues the root of the IUR-Tree from the queue and for every child E of the root, inherits the contribution list of its parent. The function $\text{UpdateCL}(E, E')$ is invoked and the contribution list of E is updated with every E' present in the candidate list, result list and the priority queue. After every invocation to $\text{UpdateCL}(\cdot)$, the algorithm checks based on the minimum and maximum bound similarity scores with the k^{th} nearest neighbor, whether to add E to the results, candidates or pruned list. If E can't be pruned or added to the results, the contribution list of E' is updated with E . This process is called the mutual effect. If E' can be added to the results or pruned, it is removed from the queue or COL .

```

32: function FINAL_VERIFICATION( $COL, PEL, Q$ )
33:   while  $COL \neq \emptyset$  do
34:     Let  $E$  be an entry in  $PEL$  with the lowest level;
35:      $PEL = PEL - \{E\}$ ;
36:     for each object  $o$  in  $COL$  do
37:       UpdateCL( $o, E$ ); //update contribution lists of  $o$ .
38:       if IsHitOrDrop( $o, Q$ )==true then // see subsection 3.3
39:          $COL = COL - \{o\}$ ;
40:       end if
41:     end for
42:     for each child node  $E'$  of  $E$  do
43:        $PEL = PEL \cup \{E'\}$ ; //access the children of  $E'$ 
44:     end for
45:   end while
46: end function

```

After updating node E with all entries of COL , ROL or U , the function *IsHitorDrop()* is again invoked. If E can't be added to the result or pruned list, a check is performed to find out whether E is a internal node or a point. If E is an internal node, it is added to the queue, else to the candidate list. When the queue becomes empty, there might be some objects left in the candidate list. The function *Final.Verification()* is invoked where the candidate objects are updated with all the entries present in PEL to decide whether they belong to result or not.

3. COUNTER-EXAMPLES

We describe three counter example in this section:

- (1) Inaccuracy regarding computation of $MinT$ and $MaxT$
- (2) Inaccuracy w.r.t. Locality Condition
- (3) Inaccuracy w.r.t. Completeness Condition

All these examples are illustrated with respect to the algorithm described in [Lu et al. 2011]; however we also explain the concepts used in constructing these examples – therefore these examples can be easily modified to suit the other algorithms. We observed that [Achtert et al. 2009] proposed an algorithm which maintains the locality condition, but violates the completeness condition. We recently observed that [Lu et al. 2014] modified their previous algorithm from [Lu et al. 2011] which now maintains the locality condition. However, their algorithm still violates the completeness condition.

3.1. Inaccuracy regarding computation of $MinT$ and $MaxT$

The branch-and-bound algorithm presented in [Lu et al. 2011] required cleverly constructed lower and upper bounds on the textual similarity (and combined textual-spatial similarity) between two groups of data objects. Its authors defined $MinT$ (minimum possible similarity) and $MaxT$ (maximum possible similarity) and claimed that these definitions, when used in conjunction with upper and lower bounds on spatial similarity, give valid upper and lower bounds on the similarity between two groups of objects. To prove this claim, they used the following crucial lemma. The first inaccuracy we report is regarding this lemma.

Definition 3.1 (Similarity Preserving Function). [Lu et al. 2011] Given two functions $fsim : V \times V \rightarrow \mathbb{R}$ and $fdim : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, where V denotes the domain of n -element vectors and \mathbb{R} , the real numbers. $fsim$ is a similarity preserving function w.r.t $fdim$,

such that for any three vectors $\vec{p} = \langle x_1 \dots, x_n \rangle$, $\vec{p}' = \langle x_1' \dots, x_n' \rangle$, $\vec{p}'' = \langle x_1'' \dots, x_n'' \rangle$, if $\forall i \in [1, n]$, $fdim(x_i, x_i') \geq fdim(x_i, x_i'')$, then we have $fsim(\vec{p}, \vec{p}') \geq fsim(\vec{p}, \vec{p}'')$.

LEMMA 3.2. [Lu et al. 2011] *Extended Jaccard is similarity preserving function wrt. function $fdim(x, x') = \frac{\min(x, x')}{\max(x, x')}$ for $x, x' > 0$.*

Counter Example. Consider three points p, p', p'' with textual vectors $\vec{p} = \langle 100, 30 \rangle$, $\vec{p}' = \langle 1, 40 \rangle$, $\vec{p}'' = \langle 1, 50 \rangle$. Using $fdim(\cdot, \cdot)$ as defined in Lemma 3.2, observe that the given points satisfy the conditions for a similarity preserving function, i.e., $\forall i \in [1, 2]$, $\frac{\min(x_i, x_i')}{\max(x_i, x_i')} \geq \frac{\min(x_i, x_i'')}{\max(x_i, x_i'')}$. However, $EJ(p, p') = 0.116 \not\geq EJ(p, p'') = 0.135$ which contradicts Definition 3.1. The *MinT* and *MaxT* formula given in the paper relied on the above Lemma to be correct, which therefore become invalid.

We now present our approach to calculate *MinT* and *MaxT* between two groups of textual objects E and E' . As explained earlier, every textual object is represented as a vector of term frequencies. For any group of objects, their intersection vector (resp. union vector) has been defined to be a vector whose every coordinate is the minimum (resp. maximum) frequency among the corresponding coordinates of objects. Denoting the intersection and union vectors of E as $\langle E.i_1, E.i_2, \dots \rangle$ and $\langle E.u_1, E.u_2, \dots \rangle$, notice that for every $o \in E$, and $j \in [1, n]$, $E.i_j \leq o.w_j \leq E.u_j$. We propose the following formulæ for *MinT*.

$$MinT(E, E') = \frac{\sum_{j=1}^n E.i_j * E'.i_j}{\sum_{j=1}^n E.u_j^2 + \sum_{j=1}^n E'.u_j^2 - \sum_{j=1}^n E.i_j * E'.i_j} \quad (3)$$

The idea for computing *MinT* is that since it is a lower bound, we want to minimize the term in the numerator and maximize the denominator of EJ to ensure that $\forall o \in E$ and $\forall o' \in E'$, $EJ(o, o') \geq MinT(E, E')$. Similarly formulæ for *MaxT* is given below:

$$MaxT(E, E') = \frac{\sum_{j=1}^n E.u_j * E'.u_j}{\sum_{j=1}^n E.i_j^2 + \sum_{j=1}^n E'.i_j^2 - \sum_{j=1}^n E.u_j * E'.u_j} \quad (4)$$

3.2. Inaccuracy w.r.t. Locality Condition

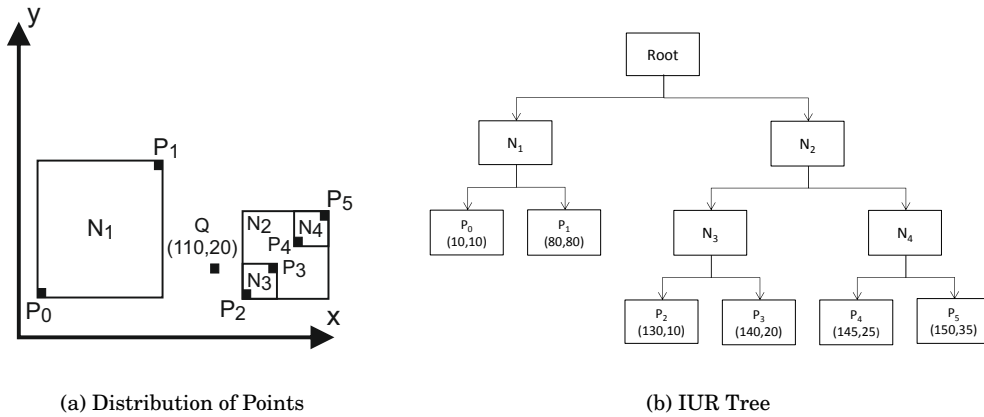


Fig. 3: Counter-example (Locality and Completeness conditions)

Consider the following counter-example for the dataset and IUR-Tree illustrated in Figure 3, and let $\alpha = 1$ and $k = 2$. The minimum and maximum distance between any two points in the database is $\varphi_s=7.07$ and $\psi_s=142.21$. The exact RST k NN of the query point Q is P_0 and P_1 . The trace of the algorithm [Lu et al. 2011] is shown in Table I. We will focus on step 1 here. The root of the tree is dequeued from the tree and node N_1 is processed. N_1 inherits the contribution lists of its parent, which is empty. Since U , ROL and COL are empty, N_1 is simply added to the queue. Now, node N_2 is processed. N_2 updates its upper and lower contribution lists with N_1 and invokes IsHitOrDrop. The upper and lower contribution lists of N_2 upon invoking IsHitOrDrop is :

$$N_2.^L.CL=\{(N_1, 0, 2)\}$$

$$N_2.^U.CL=\{(N_1, 0.68, 2)\}$$

Since $MinST(N_2, Q) = 0.73$, which is more than the upper bound given by $N_2.^U.CL$, at this point node N_2 is accepted (wrongly) as the RST k NN of Q .

Table I: Trace of RST k NN Algorithm (2011)

Steps	Actions	U	COL	ROL	PEL
1	Dequeue Root, Enqueue N_1	N_1	\emptyset	P_2, P_3, P_4, P_5	\emptyset
2	Dequeue N_1	\emptyset	\emptyset	$P_0, P_1, P_2, P_3, P_4, P_5$	\emptyset

We attribute this fault to the violation of the *Locality Condition*, a property that, we claim, must have been followed by these algorithms.

Locality Condition. Nearest neighbors of data points in a node may belong to the node itself; hence, every node should compute similarity with itself and include itself as a candidate (along with other similar nodes) in any test to prune or accept the node as RST k NN of Q .

In the counter-example above, node N_2 does not satisfy this condition since its contribution lists do not contain itself or points inside it.

3.3. Inaccuracy w.r.t. Completeness Condition

The trace of the algorithm [Lu et al. 2014] is shown in Table II.

Table II: Trace of RST k NN Algorithm (2014)

Steps	Actions	U	COL	ROL	PEL
1	Dequeue Root, Enqueue N_1 , Enqueue N_2	N_1, N_2	\emptyset	\emptyset	\emptyset
2	Dequeue N_2	N_1	\emptyset	P_2, P_3	N_4
3	Dequeue N_1	\emptyset	\emptyset	P_0, P_1, P_2, P_3	N_4

We will now focus on Step 2, when node N_2 is dequeued from the priority queue, and its children are now being processed. Node N_3 is now processed and it inherits the contribution lists of its parent N_2 . The function IsHitOrDrop is called, but N_3 can't be pruned or added to the results. After invocation of IsHitOrDrop, N_3 updates its contribution list with itself to maintain the locality condition. N_3 further updates its contribution list with other entries present in COL , ROL and U sorted in the decreasing order of the maximum spatio-textual similarity with N_3 . The upper and lower contribution list of N_3 is shown below :

$$N_3.^L.CL = \{(N_3, 0.94, 1), (N_1, 0, 2)\}$$

$$N_3.^U.CL = \{(N_3, 1, 1), (N_1, 0.68, 2)\}$$

Since $MaxST(N_3, Q) = 0.90$, which is less than 0.68; so at this point N_3 is accepted (wrongly) as $RSTkNN$ of Q . We claim that this faulty behaviour is due to not ensuring the *Completeness Condition*, viz., absence of N_4 in contribution lists of N_3 . This condition is discussed in more detail in Section 4.3. In this example, the contribution lists of N_3 is not complete.

4. PROPOSED $RSTkNN$ QUERY ALGORITHM

In this section, we present a modified algorithm to answer $RSTkNN$ queries. We will illustrate our algorithm with an example, pointing out the modifications and end this sections with a formal proof of correctness. We begin by formalizing some notions which will be used in the algorithm, and will be crucial in ensuring its correctness.

As explained earlier, the algorithms we considered worked on data that was stored in a hierarchical tree-like index, where the leaf nodes are data points themselves (to be represented by small letters) and internal nodes (to be represented by CAPITAL letters) contain pointers to children nodes. Our modified algorithm will share backbone of these algorithms; however, structurally, it will bear resemblance to the algorithm presented in [Lu et al. 2011; Lu et al. 2014]. However, it will be presented in a generalized manner which can be used to perform $RkNN$ queries, given any value of k , on a wide variety of data and independent of the explicit indexing structure used. The only requirement from the data and the index is a similarity measure $Sim(\cdot, \cdot)$ among the data points, information about the of number of objects in each node and estimates $MinSim$ and $MaxSim$ among nodes (explained below).

4.1. Contribution List a.k.a. NN -list

We will use the following notation: if e' is the k^{th} nearest neighbor of e , then we will write e' as $kNN(e)$. We will use the convention that a point is the 0^{th} nearest neighbor of itself. An immediate observation is the following: $Sim(e, kNN(e)) \geq Sim(e, k'NN(e))$ for any $k' \geq k$.

One way to answer $RkNN$ queries is by computing the list of nearest neighbors (NN -list) for every data point e : $NN(e)$ is an ordered list of data points $\langle e_1, e_2, e_3, \dots \rangle$ such that e_1 is $1NN(e)$, e_2 is $2NN(e)$ and so on. Computing this list explicitly for every data point could be very inefficient. The usual approach followed by branch-and-bound algorithms like [Achtert et al. 2009; Lu et al. 2011; Lu et al. 2014] is searching the index top-down while maintaining two NN -lists with each node - one contains an overestimate of its nearest neighbor, and another containing an underestimate of the same. These estimated lists are constructed using two functions $MinSim(\cdot, \cdot)$ and $MaxSim(\cdot, \cdot)$ which must satisfy the property below. The actual implementation of these functions depend crucially on the type of data used and the index. For two nodes E and E' ,

- $MinSim(E, E')$ must give a lower bound for the minimum similarity between pairs of points from E and E' i.e. $\forall e \in E, \forall e' \in E', Sim(e, e') \geq MinSim(E, E')$.
- $MaxSim(E, E')$ must give an upper bound for the maximum similarity between pairs of points from E and E' i.e. $\forall e \in E, \forall e' \in E', Sim(e, e') \leq MaxSim(E, E')$.

Next, we will define the main component of our algorithm, a formalization of *contribution lists* (CL) used in earlier algorithms.

Definition 4.1 (NN -list). An NN -list of a node E is a list of tuples: $\langle (E_1, m_1), (E_2, m_2) \dots \rangle$, where each E_i is a node and m_i is a positive integer.

The NN-lists we will maintain per node are $NN_U(E)$ and $NN_L(E)$ whose tuples will provide estimates to the similarity of E to its r^{th} nearest neighbor, for various values of r .

4.2. Lower bound list NN_L

The central idea behind the NN_L list comes from the following observation. Suppose for a set of m points $\{e'_1, e'_2, \dots, e'_m\}$ and another point e , we have that $Sim(e, e'_i) \geq s$. Then, it is obvious that if e does not belong to this set, $Sim(e, mNN(e)) \geq s$; and if e belongs to this set, then $Sim(e, (m-1)NN(e)) \geq s$. Extending this concept to nodes, consider any node E with m data points; now, if $MinSim(E, e) \geq s$ then, $Sim(e, mNN(e)) \geq s$ if $e \notin E$ and $Sim(e, (m-1)NN(e)) \geq s$ if $e \in E$. Notice that these bounds are tight.

We can even extend this idea to multiple nodes to get the following claim. Let e be a data point and E_1, E_2, \dots, E_k be a collection of non-overlapping nodes which do not contain e , where the list is sorted in decreasing order of $MinSim(E_i, e)$. Let m_i denote the number of data points in E_i , and let s_i be a lower bound on $MinSim(E_i, e)$. Then, for all $j = 1 \dots k$, $Sim(e, (\sum_{i=1}^j m_i)NN(e)) \geq s_j$. If $e \in E_i$ for some i , then m_i must be replaced with $m_i - 1$. We can generalize this even further by considering a node instead of e .

Definition 4.2 (Lower NN-list). An NN-list $\langle (E_1, m_1), \dots \rangle$ of non-overlapping nodes is a valid $NN_L(E)$ if:

- the list is sorted in decreasing order of $MinSim(E_i, E)$
- for all $e \in E$, if E does not overlap with E_i , then $m_i \leq |E_i|$ and if E overlaps with E_i , then $m_i \leq |E_i| - 1$

The following lemma describes the use of lower NN-lists to get underestimates of nearest neighbors. The proof is immediate from earlier definitions.

LEMMA 4.3. For any t and i that satisfies $\sum_{k=1}^{i-1} m_k < t \leq \sum_{k=1}^i m_k$ (including the case $t \leq m_1, i = 1$), it holds that for all $e \in E$, $Sim(e, tNN(e)) \geq MinSim(e, E_i)$.

4.3. Upper bound list NN_U

We want to define NN_U as an overestimation of nearest neighbors similar to NN_L and derive a similar lemma as Lemma 4.3; however, we require an additional concept first.

Definition 4.4 (Complete NN-list). We say that an NN-list $NN(E)$ is *complete* if every data point is present in some node in the NN-list, and for every (E_i, m_i) in the list,

- if E does not overlap with E_i , then $m_i = |E_i|$
- if E overlaps with E_i , then $m_i = |E_i| - 1$

It must be noted that an NN_L list need not be complete for it to satisfy Lemma 4.3. However, similar arguments do not work for NN_U . Take for example, the example situation similar to the one described for NN_L : we have a set of points $\{e'_1, e'_2, \dots, e'_m\}$ and another point e (all distinct). But even if we know that $Sim(e, e'_i) \leq s$ for some s and for all i , it is nevertheless not true that $Sim(e, mNN(e)) \leq s$, unless, all points other than e are in the set – which is precisely what a complete NN-list specifies.

Now we can define similar concepts like NN_L .

Definition 4.5 (Upper NN-list). For a node E , an NN-list $\langle (E_1, m_1), \dots \rangle$ of non-overlapping nodes is a valid $NN_U(E)$ when the following holds:

- the list is sorted in decreasing order of $MaxSim(E_i, E)$

— the list is complete

Observe that the completeness condition requires that $NN_U(E)$ must contain E itself, or its parent node, or all its children nodes – this is essentially the locality condition we mentioned earlier (Section 3.2). However, we have chosen to specifically highlight the above condition separately from the more general completeness condition. The main working lemma for NN_U follows next.

LEMMA 4.6. *For any t and i such that $\sum_{k=1}^{i-1} m_k < t \leq \sum_{k=1}^i m_k$ (including the case when $i = 1$ and $t \leq m_1$), it holds that for all $e \in E$, $\text{Sim}(e, tNN(e)) \leq \text{MaxSim}(e, E_i)$.*

4.4. Branch-and-bound traversal

A branch-and-bound algorithm traverses a hierarchical index by first visiting the root, and then exploring its children nodes, and so on. For every node it visits, the algorithm decides what to do next based on some estimate of the relevance of the current node to the desired answer (here, NN_U and NN_L lists). It may choose to further explore the node, add all the points in the node to the result set and not explore the node further (aka. *accepting* the node), or, simply not explore the node further because it decided that the node does not contain any point that should be in the result set (aka. *pruning* the node).

Suppose the query point is denoted by Q ; and suppose that a branch-and-bound algorithm is currently visiting E during its traversal of the index. Let $NN_L(E)$ denote the (valid) lower NN-list of E node, and $NN_U(E)$ denote its (valid) upper NN-list. Also, suppose i is the smallest index such that $k \leq \sum_{t=1}^i m_k$ for $NN_L(E)$, and j is the smallest similar index for $NN_U(E)$.

Here are the main theorems that give us sufficient conditions for accepting and pruning certain nodes in the index during a branch-and-bound traversal.

THEOREM 4.7 (ACCEPTING AND PRUNING CONDITION).

- (1) *If $\text{MaxSim}(E, Q) \leq \text{MinSim}(E, E_i)$, then Q cannot have any node in E in its $RkNN$ set. Therefore, E can be pruned.*
- (2) *If $\text{MinSim}(E, Q) > \text{MaxSim}(E, E_j)$, then all nodes in E belong to $RkNN$ of Q and so E can be accepted.*

The proofs for the two cases are immediate from Lemma 4.3 and 4.6, respectively. ¹

4.5. Algorithm

Now we will discuss the modified algorithm for finding reverse nearest neighbors on spatial-textual objects. Our algorithm is a modification of the one proposed in [Lu et al. 2011], so we will mostly engage in highlighting the major changes. Like the original algorithm, our algorithm uses the following data structures: a FIFO queue (U), a result list (ROL), candidate list (COL) and pruned list (PEL). We use a FIFO queue instead of a priority queue, as each entry of needs to update its NN-list with every other entry present in every list in order to ensure completeness of lists. So, the order in which other entries are added is irrelevant. We will frequently use NN-lists to refer to both the upper and lower NN-lists of the corresponding entry.

As before, the algorithm initializes the lists and enqueues the root of the IUR-tree. While the queue is not empty, an entry E is dequeued from the queue and its parent is removed from its NN-list. The two key modifications we suggest are stated next. First,

¹For accepting or pruning, in case there is a tie between similarities between query point and a database point, we tie-break in favour of points in the database. The alternative approach requires straight forward modification to the results in this subsection.

Algorithm 2 RST k NN (R : IUR-Tree root, Q : query)

```

1: Output: All objects  $o$ , s.t  $o \in \text{RST}k\text{NN}(Q, k, R)$ .
2: Initialize a FIFO queue  $U$ , and lists  $COL, ROL, PEL$ ;
3: EnQueue( $U, R$ );
4: while  $U$  is not empty do
5:    $E \leftarrow \text{DeQueue}(U)$ ; //FIFO Queue
6:   for each tuple  $\langle E'_i, num_i \rangle \in NN_L(E)$  do
7:     if  $E'_i = E$  or  $E'_i = \text{Parent}(E)$  then
8:       remove  $\langle E'_i, num_i \rangle$  from  $NN_L(E)$  and  $NN_U(E)$ ;
9:     end if
10:  end for
11:  if (then  $E$  is an internal node)
12:    Additself( $E$ ) //Ensure locality condition
13:  end if
14:  for each entry  $E'$  in  $U$  do // Ensure completeness condition
15:    Update_NN-list( $E, E'$ ); //mutual effect
16:    Update_NN-list( $E', E$ ); //mutual effect
17:  end for
18:  if  $E$  is not a hit or drop then
19:    if  $E$  is an index node then
20:      for each child  $C_E$  of  $E$  do
21:        Inherit( $NN_L(C_E), NN_L(E)$ );
22:        Inherit( $NN_U(C_E), NN_U(E)$ );
23:        EnQueue( $C_E$ )
24:      end for
25:    else
26:       $COL.append(E)$ ;
27:    end if
28:  end if
29: end while
30: Final_Verification( $COL, PEL, ROL, Q$ );

```

if E is an internal node of the tree, it adds itself to its NN-lists, thereby maintaining the *locality condition* (line 12). Then E updates its NN-lists with each entry E' present in the queue and vice versa. The updation of NN-list of E with every other entry in the queue maintains the *completeness condition* (line 14). After this, IsHitOrDrop is invoked to check if E can be pruned or added to the results. If E can neither be pruned nor added to the results, its children are added to the queue if E is an internal node; otherwise, E is added to the candidate list. We continue with the optimisation of having the children of E copy the NN-list of E before they are enqueued to U . When the queue becomes empty, there might be some candidate points left in the candidate list. The procedure Final_Verification is invoked to decide whether the points present in the candidate list belong to the result list or the pruned list; this procedure essentially checks every candidate point with other entries.

We illustrate the working of our algorithm on the example presented earlier (Figure 3) in Table III. As expected, the algorithm now correctly returns P_0 and P_1 as the only points in RST k NN of Q .

4.6. Proof of Correctness

We will now give a formal proof of correctness of our algorithm. Essentially, we will show that, when an index node is checked (line 18) if it can be immediately accepted

```

31: function FINAL_VERIFICATION(COL, PEL, ROL, Q)
32:   PEL = SubTree(PEL)
33:   while COL  $\neq \emptyset$  do
34:     for each point o in COL do
35:       for each point r in ROL do
36:         Update_NN-list(o, r);
37:       end for
38:       for ( doeach point p in PEL )
39:         Update_NN-list(o, p);
40:       end for
41:       for ( doeach point c' in COL - {o} )
42:         Update_NN-list(o, c');
43:       end for
44:       if IsHitOrDrop(o, Q)==true then
45:         COL = COL - {o};
46:       end if
47:     end for
48:   end while
49: end function

```

Table III: Trace of our algorithm

Steps	Actions	U	COL	ROL	PEL
1	Dequeue Root, Enqueue N_1 , Enqueue N_2	N_1, N_2	\emptyset	\emptyset	\emptyset
2	Dequeue N_1	N_2, P_0, P_1	\emptyset	\emptyset	\emptyset
3	Dequeue N_2	P_0, P_1, N_3, N_4	\emptyset	\emptyset	\emptyset
4	Dequeue P_0	P_1, N_3, N_4	\emptyset	P_0	\emptyset
5	Dequeue P_1	N_3, N_4	\emptyset	P_0, P_1	\emptyset
6	Dequeue N_3	N_4, P_2, P_3	\emptyset	P_0, P_1	\emptyset
7	Dequeue N_4	P_2, P_3	\emptyset	P_0, P_1	N_4
8	Dequeue P_2	P_3	P_2	P_0, P_1	N_4
9	Dequeue P_3	\emptyset	P_2	P_0, P_1	N_4, P_3
10	Verify P_2	\emptyset	\emptyset	P_0, P_1	N_4, P_3, P_2

or pruned (using Theorem 4.7), its NN-lists (especially, upper NN-list) are complete (hence, valid).

First, we want to discuss a few observations. The first fact is, if at any point of time, a data point e not belonging to an entry E is covered in $NN(E)$, then e is covered subsequently in the NN-list of E . Since e is covered at this instant, some ancestor E^* of e must be present in the NN-list of E at that instant. Observe that after an entry is added to the NN-list of E , it is removed from the NN-list of E only when the NN-list of E is updated with the children of E^* (lines 21,22). This ensures that e is forever covered in the NN-list of E .

Similarly, e is covered subsequently in the NN-lists of all (sub-)children of E . At line 18 of the algorithm, if E can't be added to the results or pruned, after updating its NN-list with each entry present in U , its children are added to the queue. However, each child of E inherits its NN-list i.e. simply copies its NN-list (lines 21,22). Therefore, the children of E will also have e in their NN-list.

Now we present the key lemma for our proof of correctness.

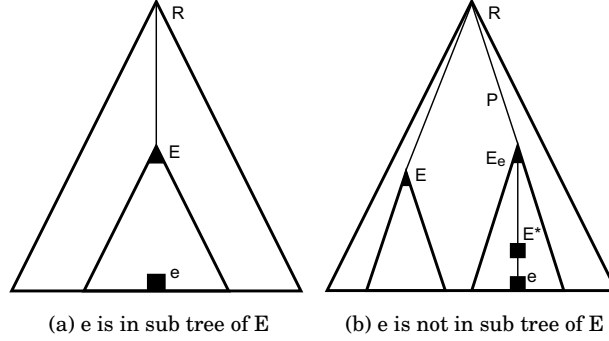


Fig. 4: Indexing tree

LEMMA 4.8. *The upper NN-list of every entry E , which is dequeued from the queue, is complete after line 17 of the RST k NN algorithm.*

PROOF. Consider an execution of the algorithm, and suppose the current node to be dequeued from the queue is denoted by E . Let e be any data point and P denote the path from root to e in the tree. We will prove that after line 17 of the algorithm, e belongs to $NN_U(E)$. There are two possibilities (see Figure 4 for reference):

- Case A. e belongs to sub tree of E
- Case B. e does not belong to sub tree of E

Case A is trivial. If e belongs to sub tree of E , it will be present in $NN_U(E)$ after line 17, since any internal node adds itself to its NN-lists (line 12).

Let us now consider Case B. Let t_1 be the time when E is dequeued from the queue. Now, one of these four different possibilities must be true at t_1 .

- Case B.1. Some node E_e on the path P belongs to the result list ROL .
- Case B.2. Some node E_e on the path P belongs to the pruned list PEL .
- Case B.3. Some node E_e on the path P belongs to the queue Q .
- Case B.4. e belongs to the candidate list COL .

Case B.1. Let t_0 denote the time when line 17 was encountered after E_e was dequeued. Once again, there are two possibilities.

Case: E belongs to the queue at t_0 . In this case, $NN_U(E)$ will contain E_e through mutual effect (line 16) at t_0 . This implies that e is covered by $NN_U(E)$ at t_1 .

Case: E does not belong to the queue at t_0 . If E does not belong to the queue, it implies that there exists some ancestor of E , say E^* (cannot be E_e because of condition of Case B.1) which belongs to the queue at time t_0 . Then $NN_U(E^*)$ contains E_e through mutual effect (line 16). This implies that once $NN_U(E^*)$ contains e , upper NN-lists of all its descendant nodes will also contain e .

The proof for Case B.2 and Case B.3 is similar to Case B.1.

We now consider the remaining Case B.4. Since $e \in COL$, it implies that some ancestor E^* of e was dequeued from the queue prior to t_1 . All the node present in the queue then contained E^* in their upper NN-list through mutual effect. Therefore at t_1 E contained E^* in its upper NN-list. \square

THEOREM 4.9. *Given an integer k , a query point Q , and an index tree R , the algorithm 2 correctly returns all $RSTkNN$ points.*

PROOF. The correctness follows from the following observations that were made earlier.

- Internal nodes are accepted or pruned (by IsHitOrDrop) only when the sufficient conditions according to the Theorem 4.7 are met (using Lemma 4.8).
- For the data points left in the candidate list COL , in Final_Verification, the (complete) NN-lists of every such point are updated with every other object (present in candidate, result and pruned list), before IsHitOrDrop being called on the point for directly accepting or pruning. Our Final_Verification routine implements this in a rather straight forward manner. In line 32 of this routine, internal nodes present in PEL are replaced with their contained points to ensure that operations in this routine directly involve points.

□

5. CONCLUSION AND FUTURE WORK

$RkNN$ is an important problem in facility location, operations research, clustering and other domains. We observed that a few published algorithms are not fully correct. In this paper we presented a correct algorithm to compute $RkNN$ on a general data set organised as a tree. We first discussed counter-examples to illustrate where the earlier algorithms made an error, and then discussed the necessity of maintaining locality and completeness conditions for ensuring the correctness of results. We finished by modifying one of the proposed algorithms along with an explanation why our algorithm is correct.

In the future, we would like to extend our algorithm for performing bichromatic $RSTkNN$ algorithm. We would further like to develop algorithms where the objects are dynamic (e.g., moving in space, or textual attributes getting updated).

REFERENCES

- Elke Achtert, Hans-Peter Kriegel, Peer Kröger, Matthias Renz, and Andreas Züfle. 2009. Reverse k-nearest neighbor search in dynamic and general metric databases. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, 886–897.
- Bhaswar B Bhattacharya and Subhas C Nandy. 2013. New variations of the maximum coverage facility location problem. *European Journal of Operational Research* 224, 3 (2013), 477–485.
- Sergio Cabello, José Miguel Díaz-Báñez, Stefan Langerman, Carlos Seara, and Inmaculada Ventura. 2010. Facility location problems in the plane based on reverse nearest neighbor queries. *European journal of operational research* 202, 1 (2010), 99–106.
- Muhammad Aamir Cheema, Wenjie Zhang, Xuemin Lin, Ying Zhang, and Xuefei Li. 2012. Continuous reverse k nearest neighbors queries in euclidean space and in spatial networks. *The VLDB Journal/The International Journal on Very Large Data Bases* 21, 1 (2012), 69–95.
- Tobias Emrich, Hans-Peter Kriegel, Nikos Mamoulis, Johannes Niedermayer, Matthias Renz, and Andreas Züfle. 2014. Reverse-Nearest Neighbor Queries on Uncertain Moving Object Trajectories. In *Database Systems for Advanced Applications*. Springer, 92–107.
- Parisa Ghaemi, Kaveh Shahabi, John P Wilson, and Farnoush Banaei-Kashani. 2012. Continuous maximal reverse nearest neighbor query on spatial networks. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 61–70.
- Antonin Guttman. 1984. *R-trees: a dynamic index structure for spatial searching*. Vol. 14. ACM.
- James M Kang, Mohamed F Mokbel, Shashi Shekhar, Tian Xia, and Donghui Zhang. 2007. Continuous evaluation of monochromatic and bichromatic reverse nearest neighbors. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 806–815.
- Flip Korn and S Muthukrishnan. 2000. Influence sets based on reverse nearest neighbor queries. In *ACM SIGMOD Record*, Vol. 29. ACM, 201–212.

- Jiahua Li, Botao Wang, and Guoren Wang. 2013. Efficient probabilistic reverse k -nearest neighbors query processing on uncertain data. In *Database Systems for Advanced Applications*. Springer, 456–471.
- Roberto J López-Sastre, Daniel Oñoro-Rubio, Pedro Gil-Jiménez, and Saturnino Maldonado-Bascón. 2012. Fast reciprocal nearest neighbors clustering. *Signal Processing* 92, 1 (2012), 270–275.
- Jiaheng Lu, Ying Lu, and Gao Cong. 2011. Reverse spatial and textual k nearest neighbor search. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 349–360.
- Ying Lu, Jiaheng Lu, Gao Cong, Wei Wu, and Cyrus Shahabi. 2014. Efficient Algorithms and Cost Models for Reverse Spatial-Keyword k -Nearest Neighbor Search. *ACM Transactions on Database Systems (TODS)* 39, 2 (2014), 13.
- Maytham Safar, Dariush Ibrahimi, and David Taniar. 2009. Voronoi-based reverse nearest neighbor query processing on spatial networks. *Multimedia systems* 15, 5 (2009), 295–308.
- Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.
- Shuo Shang, Bo Yuan, Ke Deng, Kexin Xie, and Xiaofang Zhou. 2011. Finding the most accessible locations: reverse path nearest neighbor query in road networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 181–190.
- P Tan and M Kumar Steinbach. 2011. V.(2005) Introduction to Data Mining. (2011).
- David Taniar, Maytham Safar, Quoc Thai Tran, Wenny Rahayu, and Jong Hyuk Park. 2011. Spatial network rnn queries in gis. *Comput. J.* 54, 4 (2011), 617–627.
- Quoc Thai Tran, David Taniar, and Maytham Safar. 2009. Reverse k nearest neighbor and reverse farthest neighbor search on spatial networks. In *Transactions on large-scale data-and knowledge-centered systems I*. Springer, 353–372.